UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Gašper Žejn

Diarizacija govorcev v zvočnih posnetkih

DIPLOMSKO DELO NA UNIVERZITETNEM ŠTUDIJU

Ljubljana, 2013

UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Gašper Žejn

Diarizacija govorcev v zvočnih posnetkih

DIPLOMSKO DELO NA UNIVERZITETNEM ŠTUDIJU

Mentor: prof. dr. Dušan Kodek

Ljubljana, 2013

Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Tržaška 25
1001 Ljubljana, Slovenija
telefon: 01 476 84 11
01 476 83 87
faks: 01 426 46 47
01 476 87 11
www.fri.uni-lj.si
e-mail: dekanat@fri.uni-lj.si

Št. naloge: 01905/2013
Datum: 01.03.2013

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Kandidat: **GAŠPER ŽEJN**

Naslov: **DIARIZACIJA GOVORCEV V ZVOČNIH POSNETKIH**
SPEAKER DIARIZATION OF AUDIO RECORDINGS

Vrsta naloge: Diplomsko delo univerzitetnega študija

Tematika naloge:

Diarizacija ali indeksiranje govorcev je postopek, v katerem se vhodni zvočni posnetek razdeli v segmente, ki ustrezajo identiteti govorcev. Na osnovi analize govornega signala se v postopku diarizacije ugotavljajo značilnosti govora in daje odgovor na vprašanje "kdo govori kdaj". Na osnovi pregleda obstoječih metod za diarizacijo in prosto dostopnih orodij naredite dva sistema za diarizacijo. Sistema preizkusite na več posnetkih slovenskega govora in ovrednotite njuno uspešnost.

Mentor:
prof. dr. Dušan Kodek

Fakulteta za računalništvo in informatiko
UNIVERZA V LJUBLJANI
2

Dekan:
prof. dr. Nikolaj Zimic

IZJAVA O AVTORSTVU DIPLOMSKEGA DELA

Spodaj podpisani Gašper Žejn, z vpisno številko 63030188, sem avtor diplomskega dela z naslovom:

Diarizacija govorcev v zvočnih posnetkih

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom prof. dr. Dušana Kodeka,
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela
- soglašam z javno objavo elektronske oblike diplomskega dela v zbirki "Dela FRI".

V Ljubljani, dne 20.06.13 Podpis avtorja:

Hvala Mateji in staršem za vso podporo skozi študijska leta,
mentorju za veliko umevnih nasvetov
ter vsem, ki so mi študij omogočili, vede ali nevede.

Kazalo

Seznam kratic in simbolov....iii
Povzetek....iv
Abstract....v
1 Uvod....1
2 Praktična raba za govorca specifičnih značilnosti govora....2
2.1 Verifikacija govorca....2
2.2 Identifikacija govorca....2
2.3 Diarizacija ali indeksiranje govorcev....3
2.4 Raziskovalni programi na področju prepoznave govorca....3
3 Principi in metode analize govorca....5
3.1 Sprotna in paketna diarizacija....5
3.2 Pred-poudarjanje....6
3.3 Mel-frekvenčni kepstrum koeficienti (MFCC)....7
3.4 Detekcija govora....10
3.4.1 Statistični detektor govora....11
3.5 Klasifikacija govorcev v razrede....11
3.5.1 Mešanice Gaussovih porazdelitev....13
3.5.2 Maksimizacija pričakovanja....14
3.5.3 Prikriti Markovski modeli....14
3.5.4 Bayesov informacijski kriterij....15
3.5.5 Združevanje razredov govorcev in diarizacija....16
3.6 Vrednotenje uspešnosti diarizacije....17
4 Orodja in vrednotenje....18
4.1 Orodja za diarizacijo in prepoznavo govora....18
4.1.1 LIUM SpkDiarization....18
4.1.2 SHoUT speech toolkit....19
4.2 Vzorčni posnetki....20
4.3 Vrednotenje....21
5 Rezultati....23
5.1 Vizualizacija diarizacije....26

5.1.1 Diarizacija posnetka Pogovor....26
5.1.2 Diarizacija posnetka Seja 1....27
5.1.3 Diarizacija posnetka Intervju....28
5.1.4 Diarizacija posnetka Seja 2....29
5.1.5 Diarizacija posnetka Seja 3....30
6 Sklepne ugotovitve....31
7 Literatura....33

Seznam kratic in simbolov

NIST	Nacionalni urad za standarde in tehnologijo (National Institute of Standards and Technology)
AFCP	Association Francophone de la Communication Parlée, posebna interesna skupina za Francoski govorjeni jezik.
BNSI	baza fonetično prepisanega in označenega slovenskega govora na posnetkih dnevnih novic in dnevno informativnih oddaj (ang. *Broadcast News Slovenia*)
MFCC	mel frekvenčni kepstrum koeficienti
DFT	diskretna Fourierjeva transformacija
DCT	diskretna kosinusna transformacija
SHoUT	orodje za diarizacijo SHoUT
LIUM	orodje za diarizacijo LIUM SpkDiarization
EM	algoritem maksimizacija pričakovanja
HMM	prikriti Markovski model
BIC	Bayesov informacijski kriterij
DER	metrika napake prid diarizaciji (ang. *diarization error rate*)
GPL	Splošna javna licenca (ang. *GNU General public license*)
JAR	datotečni zapis Java arhiv (ang. *Java archive*)
CMU	Carnegie Mellon University

Povzetek

Področje analize govora je v računalništvu zelo obširno področje. Diarizacija je postopek, ki išče odgovor na vprašanje "kdaj je govoril kateri govorec" z uporabo analize govora ter odkrivanja za govorca značilnih lastnosti.

Namen diplomske naloge je ovrednotiti uporabnost prosto dostopnih orodij za diarizacijo za govor v slovenskem jeziku, še posebej na govoru sestankov oziroma sej. V diplomski nalogi tako obravnavam dve orodji za diarizacijo, SHoUT in LIUM SpkDiarization. Obe orodji uporabljata podobne teoretične principe, ki so podrobneje predstavljeni v 3. poglavju. Orodji, njihova uporaba ter različni vzorčni posnetki, na katerih bomo izvajali diarizacijo, so predstavljeni v 4. poglavju.

Rezultati diarizacije kažejo, da je orodje SHoUT uspešno tudi za slovenski jezik, kljub temu, da orodje na tem jeziku v času razvoja ni bilo vrednoteno. Orodje LIUM SpkDiarization je precej manj stabilno, določene anomalije, kot je na primer združevanje govorcev enega spola v enega govorca, pa kažejo, da bi za smiselno rabo orodja na slovenskem govoru morali dodatno preučiti in nastaviti mnoge parametre.

Ključne besede

analiza govora, diarizacija, indeksiranje govorcev

Abstract

Speech analysis is a broad research area in computer science. Diarization is a process of answering the question "who spoke when" by analyzing speech and extracting speaker specific information from it.

This thesis focuses on evaluation of freely available tools for speaker diarization for use on Slovenian speech with emphasis on recordings of meetings. Two tools are evaluated, SHoUT and LIUM SpkDiarization. Both tools use similar theoretical primitives, which are explained in chapter 3. Tools, their use and test recordings are introduced in chapter 4.

Results show the SHoUT tool is useful for Slovenian speech too, despite the fact the tool was not evaluated on Slovenian speech during its development. LIUM SpkDiarization is less stable and shows peculiar anomalies, such as merging all the speakers of same gender into one, which indicates additional research and parameter discovery should be done before using LIUM SpkDiarization on Slovenian speech.

Keywords

speech analysis, diarization, diarisation, speaker indexing

1 Uvod

Obdelava in analiza človeškega govora z računalniškimi sistemi je široko in aktivno razvojno področje računalništva, ki nudi še mnogo težkih izzivov. Samodejna prepoznava govora je v prejšnjem stoletju dolgo veljala za tehnologijo, ki bo dostopna v nekaj desetletjih, a še danes se nasmihamo ob prepoznavah, kakor jih izvedejo mobilni telefoni.

Z napredkom računske zmogljivosti sodobnih procesorjev in dobro povezljivostjo mobilnih naprav je postala možna tudi prepoznava govora na strežniku. V tem primeru odjemalec pošlje govor na strežnik, ta pa vrne tekstovni zapis govora. Delujoči taki rešitvi sta Siri na telefonih znamke Apple in prepoznava govora prek strežnikov podjetja Google, ki deluje na telefonih z operacijskim sistemom Android. A ta podjetja ne dosegajo najboljših rezultatov zaradi novega razvoja v govornih tehnologijah - ta je namreč ista že zadnjih nekaj desetletij, temveč predvsem zaradi velike količine zahtevkov in posledično velike količine posnetkov, na katerih lahko ponudniki vrednotijo in učijo statistične modele. Za doseganje stabilnih algoritmov tako danes podjetja posegajo po ustvarjanju obsežnih podatkovnih baz, na katerih algoritme preverjajo. Apple in Google oba poslane posnetke zadržita za kar dve leti.

Za preizkus kakovosti diarizacije, torej ugotavljanja kdaj govori kateri govorec, je podobno potreben kvalitetno označen posnetek. Večino tovrstnih testnih posnetkov, namenjenih preverjanju analize govora (in diarizacije), so dobili z ročnim označevanjem. Včasih bi lahko tak zapis nastal že ob pravilnih nastavitvah zajema. Če bi zvočni zapis televizijskega programa hranili tudi ločeno glede na posamezen mikrofon, bi lahko v kratkem času z razmeroma malo truda dobili poleg končnega posnetka tudi zajetno količino ločenih zvočnih posnetkov, ki bi jih lahko uporabili za vrednotenje diarizacije združenega posnetka. Seveda tovrstni zajem zadostuje predvsem za diarizacijo, ne pa tudi za druge postopke pri analizi govora, na primer za prepoznavo govora.

Diarizacija ponuja možnost pridobivanja metapodatkov o govorcu tudi takrat, ko je ta informacija zaradi ene zvočne steze vsebovana le še v zvoku samem, kar odpira nove načine rabe, na primer s posnetkom obogaten prepis govora.

2 Praktična raba za govorca specifičnih značilnosti govora

Človeku je zelo naravno, da govorca prepoznamo po zvenu, barvi, višini govora in drugih značilnostih. Poskusi, da bi to delali strojno, so uspešni samo delno, z omejeno natančnostjo.

Algoritmi, ki upoštevajo za govorca specifične značilnosti, imajo kar nekaj različnih rab. V primeru, da sistem preverja identiteto, značilnosti govorca primerjamo s predhodno shranjenimi referenčnimi značilnostmi. Druga raba je iskanje identitete, kjer izmed znanih govorcev na podlagi značilnosti govora iščemo najbližja ujemanja. Tretja raba je diarizacija oz. indeksiranje govorcev, kar obravnava tudi ta diplomska naloga.

2.1 Verifikacija govorca

Verifikacija govorca je proces, ko sistem na podlagi predhodno shranjenih značilnosti govorca preverja njegovo istovetnost. Tak sistem zahteva predhoden zajem značilnosti govorca (učenje sistema), katerega identiteto se v fazi verifikacije preverja.

Kadar je verifikacija od izgovorjenih besed odvisna, gre za to, da mora govorec predhodno sistem naučiti z istimi besedami. To je uporabno kadar je slovar besed zelo omejen npr. na številke. Kadar sistem ni odvisen od izgovorjenih besed, je s strani govorca potrebno manj truda. Ker ni zahtevana izgovorjava specifičnih besed, je možen tudi pasiven vnos govorca v sistem. Velja pa omeniti, da imajo tudi lastnosti mikrofona vpliv in da se dosega boljše rezultate, če govorec za fazo učenja uporablja isti mikrofon kot za fazo verifikacije.

Pri varnostnem preverjanju je bistveno, da je verjetnost napačne verifikacije čim manjša. Značilnosti govorca se pri posamezniku razmeroma hitro spreminjajo, kar je v neposrednem konfliktu s ciljem verifikacije. Obstaja tehnična možnost, da se pri uspešni verifikaciji posodobi model govorca. Pri tem se poraja vprašanje, če je tovrstno obnašanje sistema v smislu varnosti sploh ustrezno.

2.2 Identifikacija govorca

Identifikacija govorca je proces, ko sistem iz govora neznane osebe izlušči značilnosti na podlagi katerih išče najboljše ujemanje z značilnostmi že znanih govorcev. Identifikacijo

govorca je možno izvajati tudi brez privolitve govorca, ker ne zahteva sodelovanja. Zaradi tega jo je tehnično možno uporabljati v forenzične namene npr. za oženje kroga osumljencev. Lahko pa se jo uporabi tudi za kak drug namen, kjer pa lahko potencialno prihaja do manjšanja zasebnosti.

2.3 Diarizacija ali indeksiranje govorcev

Laično rečeno nas pri indeksiranju govorcev oz. diarizaciji zanima ob katerem času v posnetku je govoril kateri govorec. Diarizacija (ang. *speaker diarization*, tudi *speaker diarisation*) je torej postopek iskanja govorcev v govoru glede na značilnosti govora posameznikov. To pomeni analizo govora in iskanje segmentov govora, kjer je govorec en, ni pa nujno, da si segmenti sledijo zaporedno. Idealni končni rezultat je skupina razredov, ki se med seboj razlikujejo po govorcu, za vsakega govorca pa je zgolj en razred z enim ali več segmenti govora. Za razliko od identifikacije govorcev, kjer iščemo znane govorce, pri diarizaciji običajno identiteta govorcev ni znana in gre torej zgolj za ločevanje različnih govorcev znotraj posnetka. Diarizacijo včasih imenujemo tudi indeksiranje govorcev (ang. *speaker indexing*).

Diarizacija se je v začetku razvila z željo, da bi informacija o govorcu pripomogla k večji natančnosti prepoznave govora. Kasneje se je zaradi uporabnosti tudi v druge namene razvoj pomembno razširil. Diarizacija je lahko dodatna informacija za nadaljnjo prepoznavo govora, lahko pa služi kot osnova za druge višje nivojske rešitve. Primer take rešitve je obogatitev prepisov sestankov z avdio (oz. video) posnetkom.

2.4 Raziskovalni programi na področju prepoznave govorca

Kar nekaj držav spodbuja razvoj govornih tehnologij z evalvacijami, na katerih preverjajo različne implementacije algoritmov in spremljajo napredek in razvoj algoritmov skozi čas.

Najpomembnejšo tovrstno evalvacijo prireja NIST, ameriški Nacionalni urad za standarde in tehnologijo. V njej ovrednotijo trenutno stanje dosežkov na področju prepoznave govora, od leta 2001 pa prirejajo tudi evalvacijo diarizacije. Zadnji krog je bil leta 2009 [1]. V okviru te evalvacije so preverjali prepoznavo govora, diarizacijo in kombinacijo obojega, torej prepoznavo govora z označbo govorcev.

Poleg razvoja orodij za angleški jezik NIST izvaja evalvacije tudi na posnetkih v arabščini in mandarinščini, najbolj razširjenem narečju kitajščine. Za druge jezike prirejajo evalvacije druge organizacije ali pa evalvacija poteka v okviru posameznih raziskav. Za francoski jezik tako evalvacije prireja AFCP[1].

Za razvoj prepoznave govora v slovenskem jeziku je bila na Univerzi v Mariboru na podlagi posnetkov dnevnih novic narejena tudi baza BNSI [3], v kateri je 36 ur govora in prepisov, ki predstavljajo učne podatke za učenje akustičnih in jezikovnih statističnih modelov za slovenski jezik.

1 Association Francophone de la Communication Parlée *je posebna interesna skupina za Francoski govorjeni jezik.*

3 Principi in metode analize govorca

Ko govorimo o obdelavi govora, je pomembno omeniti, da je informacij v govoru zelo malo, ob predpostavki, da nas zanima zgolj povedana vsebina in ne tudi npr. psihofizično stanje govorca. Prvi teoretični poskusi kompresije govora so nastali že leta 1928, ko je Homer Dudley izumil vokoder, ki pa je bil zgrajen šele leta 1939. Vokoder deluje po principu razčlembe in predstavitve govora z maloštevilnimi koeficienti amplitud spektra govora in ponovne sinteze po prenosu skozi kanal.

Podobno velja za analizo govora. Informacije, potrebne za analizo govora, so precej manj pomnilniško zahtevne, kot pa je celoten posnetek. Ker želimo obdelovati čim manj podatkov, je smiselno, da iz posnetka izvlečemo pomembne informacije.

3.1 Sprotna in paketna diarizacija

Preden pogledamo metode za obdelavo zvoka, je smiselno razčistiti razliko med sprotno in paketno obdelavo podatkov in omejitvami, ki nastanejo ob izbiri enega načina delovanja sistema.

Zaradi različnih zahtev med sprotno (ang. *on-line*) in paketno (ang. *batch*) obdelavo podatkov prihaja do razlikovanja algoritmov. Pri sprotni obdelavi je bistveno, da sistem odgovori v omejenem času. Zaradi časovne in pomnilniške omejitve so rezultati sprotnega procesiranja običajno manj natančni.

Zaradi teh dodatnih omejitev so nekateri algoritmi za uporabo pri sprotni diarizaciji neprimerni. Omejen čas odziva in omejene pomnilniške zmožnosti ne omogočajo, da bi do posnetka lahko dostopali naključno, niti ne omogočajo, da bi celoten posnetek obdelali v več prehodih. Lahko bi tudi rekli, da se pri sprotni diarizaciji osnovno vprašanje problema diarizacije ("Kdaj govori kdo?") prevesi v vprašanje "Ali je govorec še vedno isti kot je bil do sedaj?"

Sprotni sistem običajno deluje v blokih, pri čemer lahko vsak blok obdelamo po istem postopku, kot bi potekala paketna obdelava celotnega posnetka. A tak sistem je še vedno sproten in ne paketen, saj ne moremo naključno dostopati do blokov pred in za obdelovanim, četudi so del istega tekočega govora.

3.2 Pred-poudarjanje

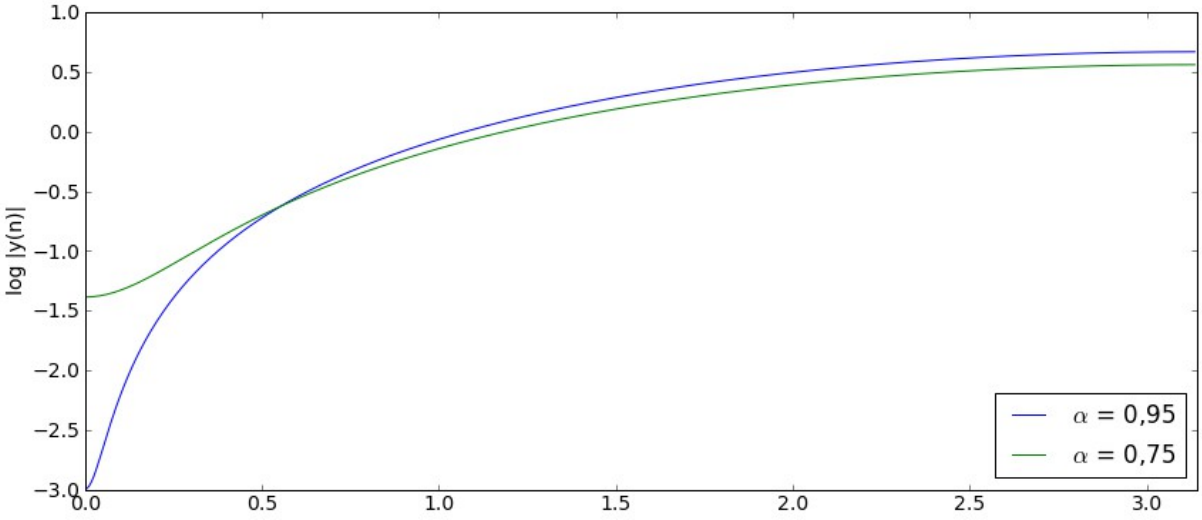
Poudarjanje je nastalo kot odgovor na to, da človeško uho različno zaznava jakost zvoka v odvisnosti od frekvence. Govor vsebuje največ informacij v frekvenčnem pasu od 2kHz do 8kHz, enako pa tudi uho najbolje zaznava frekvence v tem pasu. S poudarjanjem to frekvenčno območje na oddajnem delu ojačamo in na sprejemnem delu nazaj oslabimo. S tem dobimo boljše razmerje med signalom in šumom, ki nastane pri prenosu signala skozi medij.

Pred-poudarjanje pri procesiranju govora uporabljamo zaradi večje robustnosti. S pred-poudarjanjem ojačamo tisto frekvenčno območje signala, ki vsebujejo največ za prepoznavo govora pomembnih informacij, hkrati pa (relativno) oslabimo tiste, ki k prepoznavi ne pripomorejo.

Pred-poudarjanje (3.1) je definirano kot visokoprepustni filter s končnim enotinim odzivom. Koeficient α ima tipično vrednost blizu 1, npr. 0,95. Orodje SHoUT ima ta koeficient vdelan v program in je nastavljen na 0,97.

$$y[n] = x[n] - \alpha \cdot x[n-1] \tag{3.1}$$

Vpliv vrednosti na potek magnitudne frekvenčnega odziva je viden na sliki 1. Pomembno je opozoriti, da je mejna frekvenca prepustnega pasu odvisna od frekvence vzorčenja, zato bo tak filter pri različnih frekvencah vzorčenja prepuščal različne frekvence.

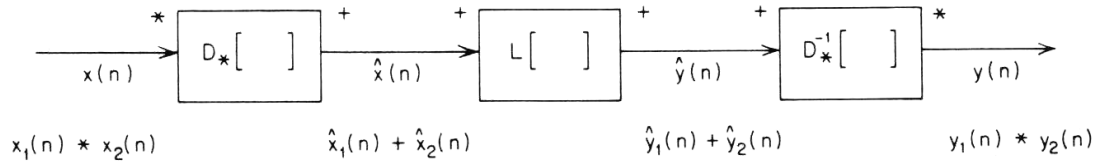


3.3 Mel-frekvenčni kepstrum koeficienti (MFCC)

Govor lahko z vidika procesiranja signalov modeliramo kot konvolucijo med vzbujanjem, ki ga dajejo glasilke in odzivom govornega trakta na enotin impulz. Vzbujanje je bodisi šum, bodisi periodičen signal. Bistveno komponento za prepoznavo govora in govorca pomeni odziv govornega trakta na enotin impulz.

Ker lahko pri prepoznavi govora opazujemo zgolj signal, bi si želeli ta signal razčleniti na vzbujanje in na odziv govornega trakta na enotin impulz. Želeli bi, da bi lahko naredili razčlembo v linearen sistem, žal pa je konvolucija v frekvenčnem prostoru produkt in ne vsota.

Transformacijam, ki pretvorijo neaditivne operacije v linearno kombinacijo, pravimo homomorfni sistemi [4]. Vse homomorfne sisteme lahko predstavimo s tremi homomorfnimi podsistemi. Prvi podsistem D_* preslika vhodno zaporedje $x(n)$, ki ni linearna kombinacija, v aditivno kombinacijo $\hat{x}(n)$, drugi sistem L je običajen linearni sistem, tretji sistem D_*^{-1} pa je inverz prvega in torej pretvori aditivno kombinacijo $\hat{y}(n)$ nazaj v neaditivno $y(n)$. Osnovni princip tega je prikazan na sliki 2.

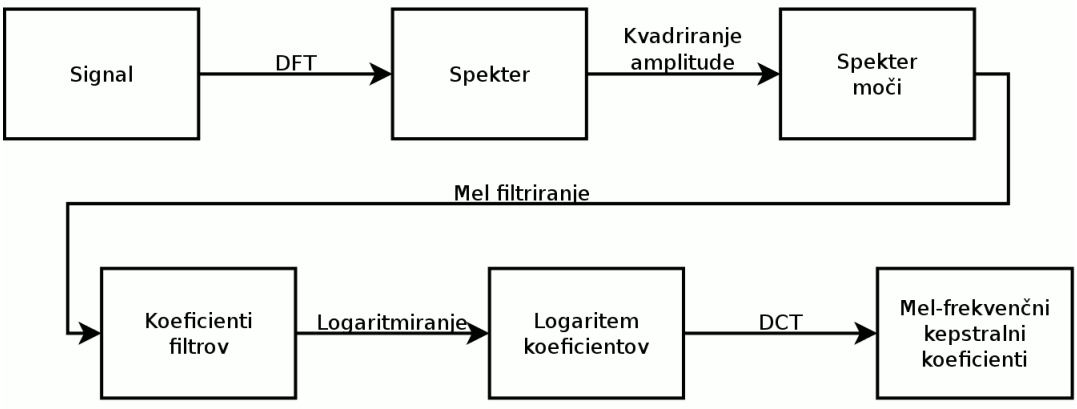


Pri analizi govora želimo dobiti koeficiente značilk, ki so medsebojno neodvisni. Spekter govora je precej redundanten, saj poleg osnovne frekvence vsebuje tudi višje harmonične frekvence. To je po eni strani zelo koristna lastnost, saj na primer ljudem omogoča dobro razumevanje govora tudi ob motnjah v posameznem frekvenčnem območju. Po drugi strani pa je tovrstna medsebojna frekvenčna odvisnost za algoritmično procesiranje nezaželena.

Beseda kepster je besedna igra na termin spekter, in sicer je zamenjan vrstni red prvih štirih črk. Z uporabo homomorfnega sistema, ki spekter preslika v kepster, želimo dobiti linearen sistem. Čeprav obstaja kar nekaj različic kepstra in podobnih (kepstralnih) transformacij, bi lahko kepster poenostavljeno definirali kot spekter spektra.

Diskretna Fourierjeva transformacija preslika periodičnost, ki se v spektru kažejo kot višje harmonične frekvence, v osnovno frekvenco, zato je uporabna tudi kot prvi del homomorfnega podsistema. Tem osnovnim frekvencam pri analizi govora pravimo formanti. Formanti so za analizo in prepoznavo govorca ali govorca zelo pomembni, ker predstavljajo značilnosti vokalnega trakta.

Postopek za pridobivanje mel-frekvenčnih kepstralnih koeficientov je prikazan na sliki 3. Nad signalom najprej izvedemo diskretno Fourierjevo transformacijo, da dobimo spekter. V drugem koraku absolutno vrednost spektra kvadriramo, da dobimo spekter moči. Nad spektrom moči v tretjem koraku izvedemo skupino mel filtrov, s tem pa dobimo zaporedje koeficientov. Te koeficiente v četrtem koraku logaritmiramo, da dobimo jakost, kakor jo zaznava človeško uho. Zadnji korak je diskretna kosinusna transformacija, s katero harmonične frekvence preslikamo v osnovne in tako dobimo mel-frekvenčne kepstralne koeficiente. Prvi koeficient v zaporedju se pogosto ne smatra kot del MFCC, saj predstavlja enosmerno komponento (DC) oz. logaritem energije.

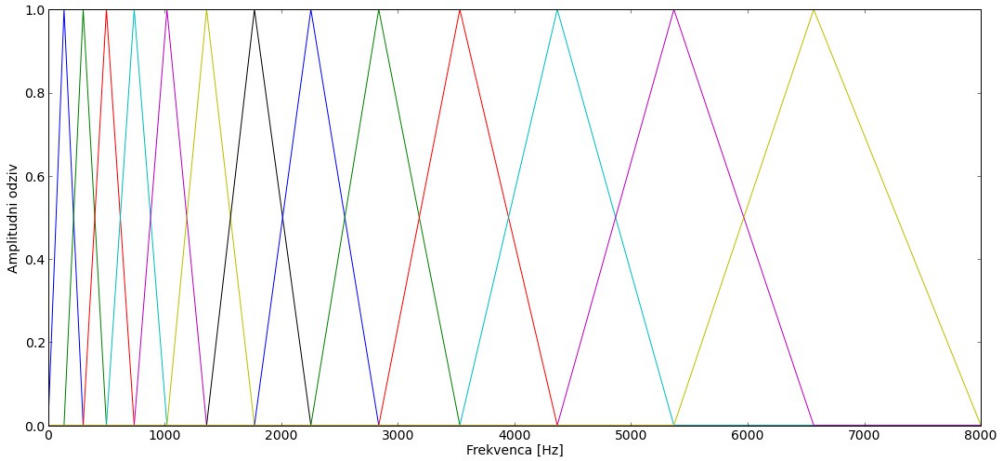


Poleg koeficientov MFCC se za potrebe analize pogosto vzame še hitrost spreminjanja koeficientov oz. delta vrednosti, ΔMFCC in včasih tudi delta-delta vrednosti ΔΔMFCC. Za ΔMFCC so to kar razlike med N-tim predhodnim in N-tim sledečim vektorjem MFCC koeficientov, za ΔΔMFCC pa razlike med ΔMFCC.

Mel filtri so trikotni filtri, ki imajo prepustne pasove določene tako, da so prilagojeni zaznavanju človeškega ušesa. Do 1000Hz so frekvenčni pasovi filtrov ekvidistančni, nad tem pa so vedno širši, tako da ustrezajo zaznavanju. Enota *mel* je namenjena označbi višine tonov. Pretvorba iz frekvenčnih enot v mel je določena z enačbo

$$mel = 1127{,}01048 \cdot \ln\left(1 + \frac{f}{700}\right) \quad . \tag{3.2}$$

Pri tem se poraja vprašanje kolikšno je ustrezno število filtrov. Pri prepoznavi govora se običajno uporablja tipično 13 filtrov, pri prepoznavi govorca oz. diarizaciji pa je bolj običajno število filtrov okrog 20. Pri manjšem številu filtrov značilke povprečijo razlike med govorci, zato je to bolj primerno za prepoznavo govora. Pri večjem številu filtrov so razlikovanja med govorci zaradi manj povprečenja v značilkah bolj izrazita, zato je to bolj primerno za iskanje razlik med govorci in diarizacijo. Orodje SHoUT uporablja za diarizacijo 20 filtrov, medtem ko orodje LIUM uporablja 13 filtrov.

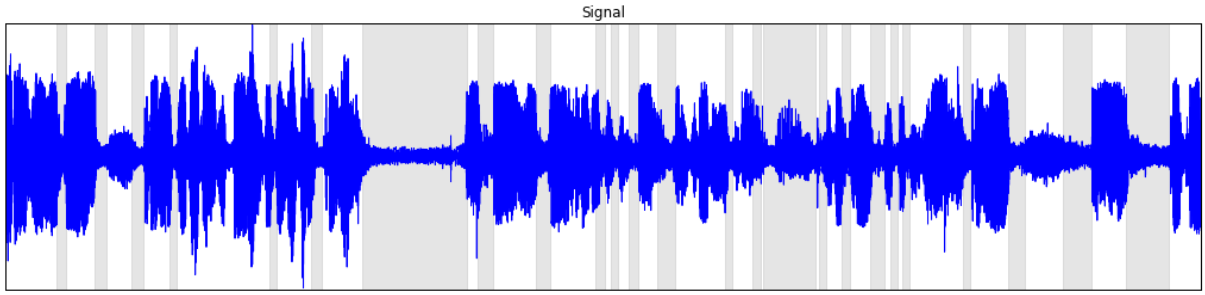


MFCC koeficienti so kot značilke osnova za vse sodobne raziskave na področju prepoznave govora in identifikacije govorcev.

3.4 Detekcija govora

Pomemben del procesiranja je razločevanje tišine, govora in negovora. Iz modelov govorcev želimo izločiti šum in druge negovorni zvok, ker z izključitvijo tišine in negovornih zvokov (npr. glasbe) precej pridobimo na natančnosti, poleg tega pa prihranimo tudi na procesni zahtevnosti. Angleški termin za detekcijo govora je *speech activity detection* ali *voice activity detection*.

Obstaja več načinov detekcije govora oz. tišine. Najenostavnejši je verjetno opazovanje energije signala. Z lastnimi empiričnimi poizkusi sem na danih posnetkih dobre rezultate dosegal s predpostavko, da je tišina, kadar je logaritmirana energija signala pod 40 percentilom energije v celotnem posnetku.



Nekoliko naprednejša detekcija govora je običajno sestavljena iz algoritmov, ki zaznajo frekvence govora. Šele ob pogoju, da so prisotne frekvence govora, se upošteva, da je signal zares govor.

3.4.1 Statistični detektor govora

Primer danes zelo dobrega detektorja govora je detektor, vključen v orodje za prepoznavo govora SHoUT. Razlikuje med tišino, govorom in med negovornim zvokom. Klasifikacija deluje s hipotezami, da ima najmanj energije tišina, negovorni zvok največ, govor pa srednje.

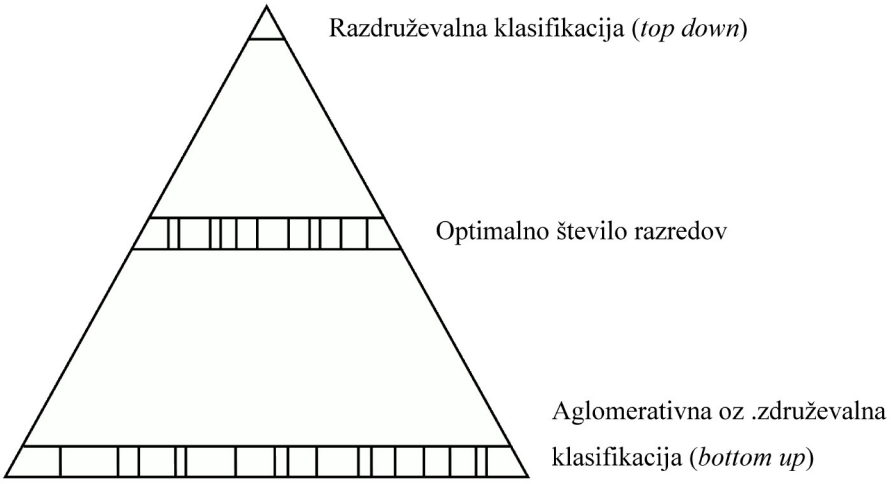
Detektor govora je osnovan na prikritih Markovskih modelih in zahteva zelo malo predznanja o vrsti zvoka, ki ga bo prepoznaval. Algoritem se z iterativnim postopkom z dvema korakoma nauči razločiti tišino od zvoka. V prvem koraku se sistem nauči modela tišine, v drugem koraku pa se ta model uporabi na podatkih za novo razdelitev na tišino in zvok. Za izvedbo tega postopka je potreben začetni model (ang. *bootstrap*) tišine in zvoka. Za uporabo je dostopen model, ki je bil naučen na nizozemskem govoru. Uporaba tega modela na angleškem govoru je pokazala, da jezik govora ne vpliva bistveno na rezultate detekcije govora [1].

3.5 Klasifikacija govorcev v razrede

S klasifikacijo značilk v razrede želimo ustvariti po eno razred za vsakega govorca. Uporablja se več vrst klasifikacije, predvsem bomo izpostavili zaporedno in hierarhično klasifikacijo.

Pri zaporedni klasifikaciji izhajamo iz predpostavke, da dva sosednja bloka značilk pripadata bodisi enemu govorcu v primeru da ni prišlo do spremembe govorca, bodisi dvema govorcema, v kolikor je do spremembe govorca prišlo. Z zaporedno klasifikacijo lahko dobimo večje začetne razrede in s tem prihranimo nekaj procesne zahtevnosti, saj ni potrebna primerjava časovno zelo oddaljenih značilk.

Glede na pristop obstajata dve vrsti hierarhične klasifikacije: združevalno oz. aglomerativno (ang. *agglomerative*, *bottom up*) in razdruževalno (ang. *divisive*, *top down*). Pri združevalni klasifikaciji je začetno izhodišče veliko število razredov, ki jih na podlagi ocene podobnosti združujemo v manjše število razredov, dokler ne dosežemo optimalnega števila, ki ustreza številu govorcev. Pri razdruževalni klasifikaciji je začetno stanje en razred, ki ga cepimo na podrazrede, dokler enako kot prej ne dobimo optimalnega števila razredov.



Slika 6: Ponazoritev aglomerativnega in razdruževalnega pristopa h klasifikaciji v razrede

Kadar je število dobljenih razredov manjše od dejanskega števila govorcev pravimo, da je prišlo do prekomernega združevanja razredov (ang. *over-clustering*), kadar pa je število dobljenih razredov večje od dejanskega števila govorcev, pa je prišlo do premalo združevanja razredov (ang. *under-clustering*). Oboje predstavlja odmik od optimalnega števila razredov in zato prispeva k večji napaki diarizacije.

V teoriji sta oba pristopa, tako aglomerativni kot razdruževalni, enakovredna in privedeta do istega rezultata. Zaradi lažje implementacije in razumevanja pa v orodjih prevladuje

uporaba aglomerativnega pristopa. Tako aglomerativni pristop uporabljata orodje SHoUT kot tudi orodje LIUM.

3.5.1 Mešanice Gaussovih porazdelitev

Mešanice Gaussovih porazdelitev so statistični model za predstavitev porazdelitve kot vsoto več Gaussovih porazdelitev [5]. Mešanice Gaussovih porazdelitev so uporabne, ko je porazdelitev sestavljena iz več porazdelitev iste vrste z različnimi parametri. Pri prepoznavi govorca je model mešanih normalnih porazdelitev uporaben, ker lahko značilke obravnavamo kot D dimenzionalni vzorec, ki so za vsakega govorca po dimenzijah porazdeljene po normalni porazdelitvi.

Normalno porazdelitev za D dimenzij zapišemo kot

$$N(x|\mu,\Sigma)=\frac{1}{(2\pi)^{D/2}}\cdot\frac{1}{|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\} \quad , \tag{3.3}$$

pri čemer so x vektorji značilk, μ D-dimenzionalni vektor srednjih vrednosti, Σ je kovariančna matrika, z $|\Sigma|$ pa označimo determinanto kovariančne matrike Σ.

Kadar z eno Gaussovo porazdelitvijo ne moremo opisati kompleksne distribucije nepovezanih spremenljivk, lahko uporabimo mešanico Gaussovih porazdelitev. Mešanice so verjetnostna vsota $p(x)$ posameznih porazdelitev z različnimi lastnostmi

$$p(x)=\sum_{k=1}^{K}\pi_k N(x|\mu_k,\Sigma_k) \quad . \tag{3.4}$$

Če imamo K Gaussovih porazdelitev, morajo biti koeficienti mešanja π na intervalu [0, 1], njihova vsota pa mora biti 1, saj mora biti skupna verjetnost enaka gotovosti

$$\sum_{k=1}^{K} \pi_k = 1 \, ; \pi_k \in [0,1] \quad . \tag{3.5}$$

Zaradi koeficientov mešanja lahko pripadnost vzorcev razredu obravnavamo kot novo skrito spremenljivko *z*, mešanico Gaussovih porazdelitev pa lahko opišemo kot vsoto pogojnih verjetnosti, da vzorec *x* pripada razredu *z*

$$p(x) = \sum_{z} p(z)\, p(x|z) = \sum_{k=1}^{K} \pi_k \, N(x|\mu_k, \Sigma_k) \quad . \tag{3.6}$$

Pri klasifikaciji v razrede nas zanima verjetnost, da vzorec pripada razredu z_k.

3.5.2 Maksimizacija pričakovanja

Da bi veliko število vzorcev uvrstili v razrede, moramo nekako oceniti parametre μ, Σ in π posameznih Gaussovih porazdelitev v mešanici, s katero razrede opisujemo. Na teoretični ravni to storimo tako, da maksimiziramo logaritmirano funkcijo verjetja (ang. *log maximum likelihood*). Za rešitev tega problema se uporablja algoritem maksimizacija pričakovanja (ang. *expectation maximization*, *EM*).

Maksimizacija pričakovanja je iterativni algoritem. Vsaka iteracija poteka v dveh korakih, ki se imenujeta E (iz *expectation*) in M (iz *maximization*). V koraku E računamo verjetnost pripadnosti vzorcev posameznim Gaussovim porazdelitvam (skrite spremenljivke z_k) ob nespremenljivih parametrih μ, Σ in π, v koraku M pa izračunamo nove parametre na podlagi skritih spremenljivk. Pogoj za končanje zanke je konvergenca parametrov ali logaritma funkcije verjetja.

3.5.3 Prikriti Markovski modeli

Klasifikacija v razrede poteka z uporabo mešanice Gaussovih porazdelitev, pri čemer pa se ne upošteva časovna komponenta govora oz. sosledje vzorcev. Ker je časovno sosledje pomemben aspekt govora, se za analizo in klasifikacijo govorca uporabljajo prikriti Markovski modeli (ang. *hidden Markov models*, *HMM*).

Markovski model je verjetnostni model, katerega porazdelitev prehodov v naslednje stanje je odvisna zgolj trenutnega stanja sistema. Tak sistem torej nima pomnjenja, njegova stanja pa so vidna, saj so obenem tudi izhodne spremenljivke.

Prikriti Markovski model je Markovski model, katerega stanje je ločeno od izhodnih spremenljivk in je prikrito. Na stanje sistema lahko sklepamo zgolj iz opazovanja izhodnih spremenljivk. Te spremenljivke so s stanjem povezane, običajno pa ne nudijo dovolj informacij, da bi lahko glede na njih konkretno določili stanje oz. pot skozi sistem. Zaradi teh lastnosti take modele imenujemo prikrite. Ker so stanja in izhodne spremenljivke nepovezane, so taki modeli precej močnejši v smislu opisovanja sistemov, obenem pa so računsko razmeroma nezahtevni, če jih primerjamo z modeli, ki bi bili odvisni tudi od prejšnjih stanj.

Zaradi njihove moči opisovanja modelov so prikriti Markovski modeli v široki uporabi, med drugim tudi za npr. analizo jezikov, prepoznavo pisave in pa seveda za prepoznavo govora in govorca.

3.5.4 Bayesov informacijski kriterij

Ker bo razredov pri združevalni hierarhični klasifikaciji več, kot je govorcev, jih želimo združiti po pripadnosti govorcu. Pri tem je prvi problem iskanje dveh razredov, ki ju želimo združiti. Drugi problem je ovrednotenje koristnosti združevanja razredov. Oba problema lahko rešimo z Bayesovim informacijskim kriterijem (ang. *Bayes information criterion*, *BIC*).

BIC je metrika prileganja statističnega modela podatkom. Je ocena (približek), ki pozitivno vrednoti prileganje modela podatkom, poleg tega pa negativno vrednoti kompleksnost modela. Kompleksnost modela se meri v številu parametrov, ki jih ima statistični model. Če je S množica N vzorcev, ki jih modeliramo s statističnim modelom M, ki ima $num(M)$ parametrov, je BIC definiran kot

$$BIC(M) = \log L(S, M) - \frac{1}{2} num(M) \log N \quad . \tag{3.7}$$

BIC je pozitiven, kadar se model dobro prilega podatkom, ali negativen, kadar se model podatkom ne prilega najbolje. Ovrednotenje koristnosti združevanja razredov lahko rešimo z

BIC, pri čemer se pri združevanju razredov odločamo ali se danim podatkom bolje prilegata dva ločena modela ali en sam.

Razliki med vrednostjo BIC enotnega modela in BIC dveh posameznih modelov rečemo tudi delta BIC

$$\Delta BIC(M_a, M_b) = BIC(M_{a\cup b}) - (BIC(M_a) + BIC(M_b)) \quad . \tag{3.8}$$

Kadar je ΔBIC pozitiven je združevanje razredov smiselno. Ob tem vidimo, da lahko ΔBIC uporabimo tudi za mero podobnosti razredov. Dokler obstajata dva razreda, katerih ΔBIC je pozitiven, ju je glede na ta kriterij smiselno združiti. Zaustavitveni pogoj za postopek združevanja razredov pa lahko definiramo kot pogoj, da ni več dveh razredov, ki bi ju bilo smiselno združiti.

3.5.5 Združevanje razredov govorcev in diarizacija

Pri klasifikaciji mel-frekvenčnih značilk v razrede, ki predstavljajo posameznega govorca, se prikriti Markovski model uporablja za segmentacijo govora po govorcih. V grobem je potek segmentacije sledeč.

1. V prvem koraku z uporabo mešanic Gaussovih porazdelitev opravimo klasifikacijo v razrede, ki predstavljajo po enega govorca.
2. V drugem koraku na teh razredih naučimo prikriti Markovski model, katerega izhodne vrednosti predstavljajo pripadnost vzorca razredu oz. govorcu. Z uporabo dobljenega prikritega Markovskega modela opravimo ponovno segmentacijo posnetka na govorce. Dobimo nove razrede značilk po pripadnosti govorcu, na katerih lahko naučimo nov prikrit Markovski model. Učenje prikritega markovskega modela in segmentacijo ponovimo trikrat.
3. V tretjem koraku združujemo razrede na podlagi podobnosti, dokler je združevanje koristno po Bayesovem informacijskem kriteriju.
4. V zadnjem koraku ponovno naučimo prikriti Markovski model in še zadnjič segmentiramo posnetek govora, da dobimo končno razdelitev govorcev.

3.6 Vrednotenje uspešnosti diarizacije

Za ocenjevanje uspešnosti diarizacije se je uveljavila metrika DER (ang. *diarization error rate*), ki je bila razvita za evalvacije ameriškega Nacionalnega urada za standarde in tehnologijo.

Ker orodja za diarizacijo ne poznajo identitete govorcev, bo praviloma vsako orodje posameznim govorcem dodelilo poimenovanja, ki se bodo razlikovala od referenčnih. Zaradi tega je prvi korak pri vrednotenju iskanje optimalne preslikave med referenčnimi poimenovanji govorcev in poimenovanji, kakor jih je dodelilo orodje za diarizacijo. Za optimalno preslikavo velja tista, ki daje najmanjšo napako po spodaj opisanem kriteriju DER.

Imamo množico S vseh odsekov posnetka. Znotraj nobenega odseka s ni zabeležena sprememba govorca niti v referenčni diarizaciji niti v diarizaciji, dobljeni z orodjem, ki ga vrednotimo. DER izračunamo po enačbi

$$DER = \frac{\sum_{s=1}^{S} dur(s) \cdot (max(N_{ref}(s), N_{sys}(s)) - N_{correct}(s))}{\sum_{s=1}^{S} dur(s) \cdot N_{ref}} \quad . \tag{3.9}$$

Pri tem je $dur(s)$ dolžina odseka s v sekundah, N_{ref} je število govorcev v referenčni diarizaciji, N_{sys} pa število govorcev, ki jih je zaznalo orodje, ki ga vrednotimo. $N_{correct}$ je število oznak govorcev, ki se v diarizaciji orodja ujemajo z referenčno, in prestavlja število pravilno označenih govorcev. Ob popolnoma pravilni diarizaciji bo metrika DER enaka 0.

Pravilnih razrezov posnetka je lahko več, hkrati pa so lahko še vedno pravilni v smislu človeškega zaznavanja. Zaradi tega je NIST definiral interval 250 ms, znotraj katerega je lahko prehod govorca, da še vedno velja za pravilnega.

Za vrednotenje diarizacije smo uporabili skripto *md-eval-21.pl*, ki jo uporablja tudi NIST.

4 Orodja in vrednotenje

4.1 Orodja za diarizacijo in prepoznavo govora

Za praktično ovrednotenje smo uporabili dve orodji za diarizacijo, ki sta bili obe razviti v okviru akademskih raziskovalnih projektov in ovrednoteni na evalvacijah, ki jih je prirejal NIST. Obe orodji uporabljata iste principe, ki so bili opisani v poglavju 3, se pa razlikujeta v načinih kombinacije metod in posameznih parametrih.

Obe orodji podpirata večnitno procesiranje in znata do izkoriščati sodobne večjedrne procesorje. Obe orodji sta na voljo pod odprtokodno licenco GNU GPL različice 2[2].

4.1.1 LIUM SpkDiarization

Orodje LIUM SpkDiarization je spisano v programskem jeziku Java in je bilo razvito pod okriljem Laboratorija za informatiko Univerze v Mainu (LIUM). Je naslednik oziroma reimplementacija orodja mClust, ki je bilo spisano v programskem jeziku C++. LIUM SpkDiarization je na voljo na spletni strani LIUM v obliki arhiva JAR [8]. Zaradi enostavnosti bomo orodje LIUM SpkDiarization v nadaljnjem besedilu imenovali LIUM. Poženemo ga z ukazom

```
java -Xmx2048m -jar LIUM_SpkDiarization-3.1.jar --fInputMask=datoteka_%s.wav
    --sOutputMask=datoteka_%s.dia datoteka.wav
```

Diarizacija se izvede v več ločenih podpostopkih, ki za vhode vzamejo rezultate prejšnjih postopkov. Diarizacijo je mogoče izvesti tudi z zaporednim klicem posameznih podpostopkov. Ker posamezni podpostopek sprejme vrsto parametrov, obširno opisovanje vsakega ni smiselno, bomo pa povzeli osnovni princip delovanja ukazov.

Prvi postopek je segmentacija. Ukaz *MSegInit* izvede preverjanje, da datoteka z značilkami MFCC ne vsebuje dveh zaporednih identičnih vektorjev značilk, zaradi katerih lahko pride do težav pri segmentaciji. Z ukazom *MSeg* pa izvedemo segmentacijo na podlagi razdalje med segmenti.

2 Besedilo GNU GPL 2.0 je dostopno na spletu na naslovu http://www.gnu.org/licenses/gpl-2.0.html.

Drugi postopek je združevanje segmentov v razrede. Ukaz *MClust* je namenjen združevanju v razrede na podlagi metrike BIC. Omogoča več načinov združevanja in se običajno uporablja v dveh korakih: v prvem koraku se združuje zaporedne segmente od leve proti desni glede na časovno zaporedje značilk, pri čemer izkoriščamo časovno lokalnost govorca, v drugem pa se tako dobljene večje segmente združuje hierarhično.

Tretji postopek je učenje prikritega Markovskega modela in ponovna segmentacija. Z ukazom *MTrainInit* najprej inicializiramo podatkovne strukture za prikriti Markovski model, z ukazom *MTrainEM* pa naučimo modele z EM algoritmom. Z ukazom *MDecode* lahko naučene modele uporabimo za ponovno segmentacijo posnetka na podlagi prikritega Markovskega modela, s čimer izboljšamo natančnost segmentacije.

Ker je orodje spisano v Javi, so se avtorji odločili, da kot knjižnico vključijo tudi CMU Sphinx, ki ga konkretno uporabljajo za pretvorbo posnetka v značilke. Zaradi istega programskega jezika pa je seveda enostavna tudi integracija LIUM v Sphinx.

Izhodna datoteka vsebuje diarizacijo v obliki metapodatkov, ki odsekom posnetka dodelijo govorca.

4.1.2 SHoUT speech toolkit

Orodje SHoUT je spisano v programskem jeziku C++. Spisano je bilo v okviru doktorske disertacije Marijna Huijbregtsa na Univerzi v Twente [6]. Za vrednotenje sem uporabil zadnjo različico 0.3, izdano 1. decembra 2010.

Avtorja orodja SHoUT je v razvoju vodila želja po čim manj nastavljivih parametrih in empirično dobljenih in statično nastavljenih (magičnih) vrednostih, ki bi implicitno predstavljala predznanje sistema o posnetkih, nad katerimi se diarizacija izvaja. Zaradi tega je sam postopek diarizacije z orodjem SHoUT zelo poenostavljen.

Diarizacijo s SHoUT orodjem izvedemo v dveh korakih. Prvi korak je detekcija govora, tišine in negovornega zvoka. Izvedemo jo z

```
shout_segment --audio posnetek.wav --am-segment shout.sad --meta-out posnetek.sad
```

Datoteka *shout.sad* vsebuje začetni statistični model za detekcijo govora, ki je bil naučen na podatkih v nizozemščini. Detektor govora deluje po postopku, ki je opisan v poglavju 3.4.1.

Drugi korak je diarizacija, ki jo izvedemo z ukazom

```
src/shout_cluster --audio posnetek.wav --meta-in posnetek.sad --meta-out posnetek.dia
```

Za vhodna parametra vzame posnetek in razdelitev na govor in negovor, dobljeno z detekcijo govora, izhod pa je diarizacija v obliki metapodatkov, ki označujejo segmente z govorci.

4.2 Vzorčni posnetki

Za vrednotenje orodij za diarizacijo sem uporabil več različnih posnetkov. Prvi posnetek je posnetek pogovora, v katerem se govorci zelo veliko prekrivajo, dodatno pa je posnetek nastal na prostem, tako da se v ozadju slišijo tudi šumi iz okolice. Trije posnetki so zvočni posnetki sej Državnega zbora Republike Slovenije. Značilnost vseh sej je, da predsedujoči daje besedo. Prvi posnetek ima malo govorcev in zelo dolge govore, posledično pa tudi malo prehodov med govorci. Posnetek druge seje ima srednje število prehodov med govorci, posnetek tretje seje pa ima več prehodov med govorci. Peti posnetek je zvočni posnetek TV oddaje Intervju, kjer sta govorca samo dva, a se interaktivno izmenjujeta, zaradi česar je prehodov govorcev zelo veliko. V spodnji tabeli so predstavljene tudi nekatere značilnosti posnetkov.

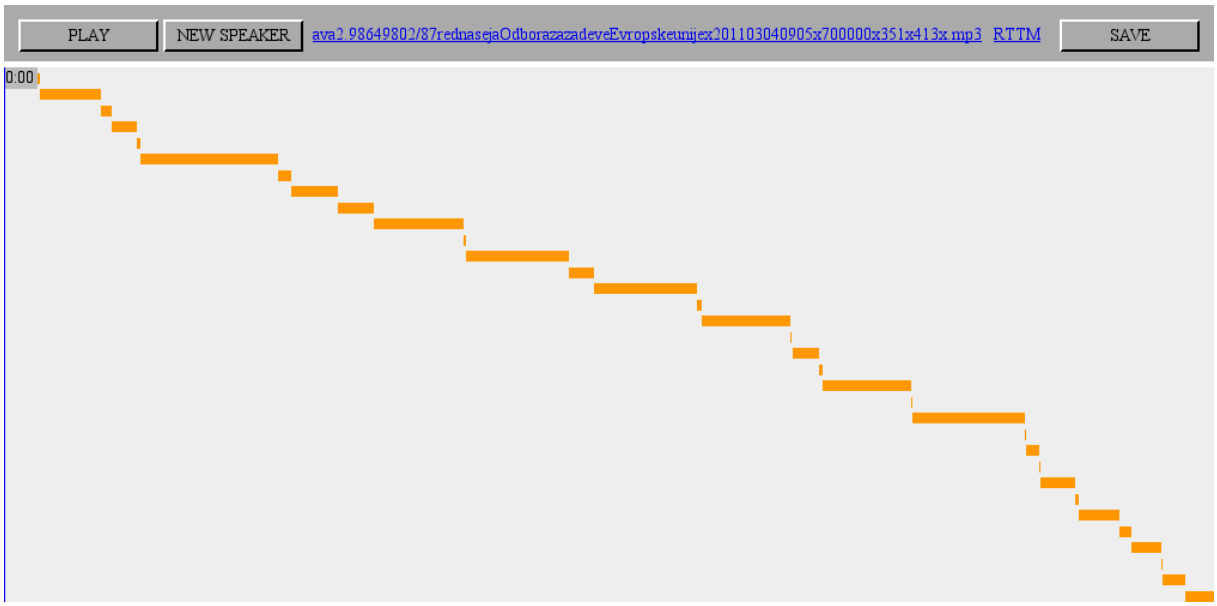
	Oznaka posnetka	**Dolžina posnetka [s]**	**Refrenčno število govorcev**	**Referenčno število prehodov govorcev**
1.	Pogovor	343,22	3	139
2.	Seja 1	1842,34	5	9
3.	Intervju	3002,85	2	236
4.	Seja 2	3147,67	11	33
5.	Seja 3	4101,36	10	48

Tabela 1: Lastnosti vzorčnih posnetkov. Posnetki so razporejeni po dolžini trajanja.

Posnetek pogovora je terenski posnetek pogovora, del zbirke digitalnega arhiva Glasbenonarodopisnega inštituta ZRC SAZU.

Posnetek intervjuja in posnetki sej pa so bili preneseni s spletne strani Radiotelevizije Slovenija, ki poleg produkcije oddaje Intervju skrbi tudi za video prenos parlamentarnih sej.

Referenčno diarizacijo sem pri posnetkih sej pridobil z ročnim označevanjem posnetkov. Za označevanje sej sem uporabil enostavno spletno aplikacijo, ki sem jo izdelal sam. Omogoča enostavno označevanje prehodov govorca. Za označevanje sej, kjer so časi med menjavo govorca daljši, je tovrstno orodje delovalo dobro, za Intervju in za posnetek pogovora, kjer se govorci izmenjujejo zelo hitro, nekoliko pa se govor celo prekriva, pa to zaradi kratkih odsekov in minimalnih premorov med govorcema ni ustrezalo. Posnetek intervjuja sem zato označil z uporabo urejevalnika glasbe Audacity in uporabo steze z označbami.



4.3 Vrednotenje

Obe orodji za diarizacijo kot izhodni podatek dajeta segmentacijo posnetka. Ta segmentacija predvideva, da imamo za raziskovalne namene bazo označenih posnetkov, v katerih imamo ločeno označene govor, tišino in druge zvoke. Ker sem sam označeval le

spremembe govorca in ker za potencialno praktično rabo to ni obvezno, sem kriterije ustreznosti diarizacije nekoliko omilil.

Vrednotenje sem izvedel tako, da sem zaporedne segmente, ki so ločeni s tišino in ki pripadajo istemu govorcu, združil v en segment. Za potrebe razvoja orodja za diarizacijo to ni ustrezno, za ovrednotenje za potencialno rabo v praksi, kjer nam informacija o govorcu prinese veliko informacij, pa tako fina razdelitev ni potrebna in ustreza tudi groba.

Kot orodje za vrednotenje sem uporabil skripto *md-eval-v21.pl* [7], spisano v programskem jeziku Perl, ki jo uporabljajo tudi za vrednotenje na evalvacijah NIST. Poleg metrike DER sem primerjal še nekatere druge parametre: število govorcev v dobljeni diarizaciji in čas za izvedbo diarizacije.

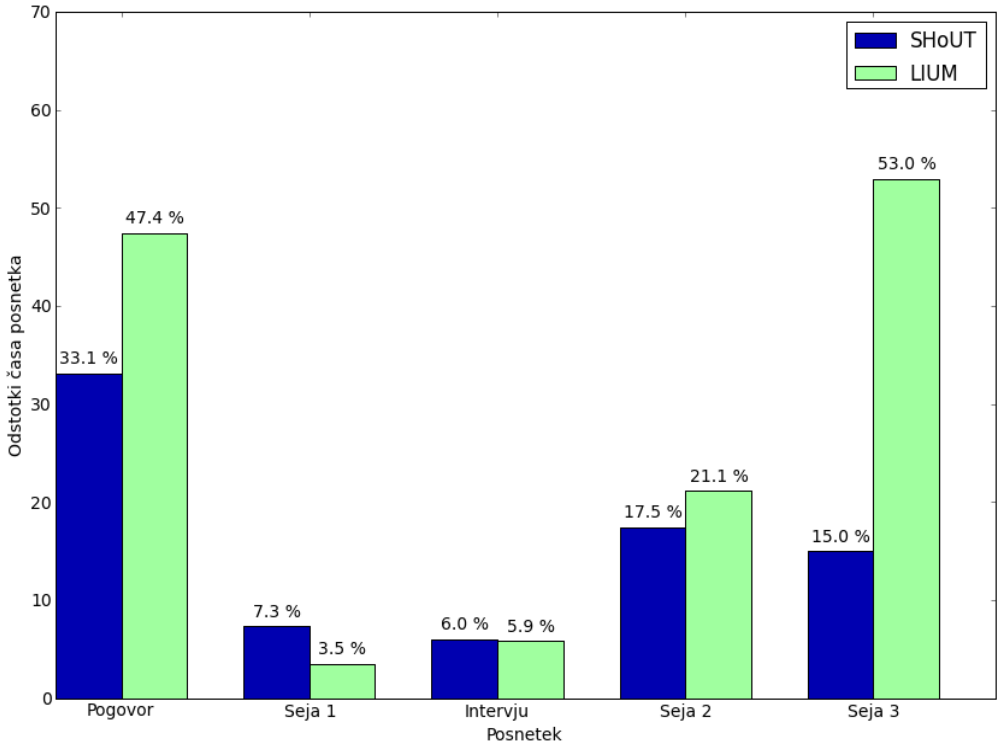
Diarizacijo sem izvajal na neobremenjenem prenosnem računalniku s štirijedrnim 2,4GHz procesorjem Intel i5 520M z 8 GB pomnilnika, ki je bil priklopljen na omrežno napetost. Obe orodji sta večnitni, tako da sta pri procesiranju izkoriščali več jeder.

5 Rezultati

Rezultati diarizacije na vzorčnih posnetkih so prikazani v spodnji tabeli.

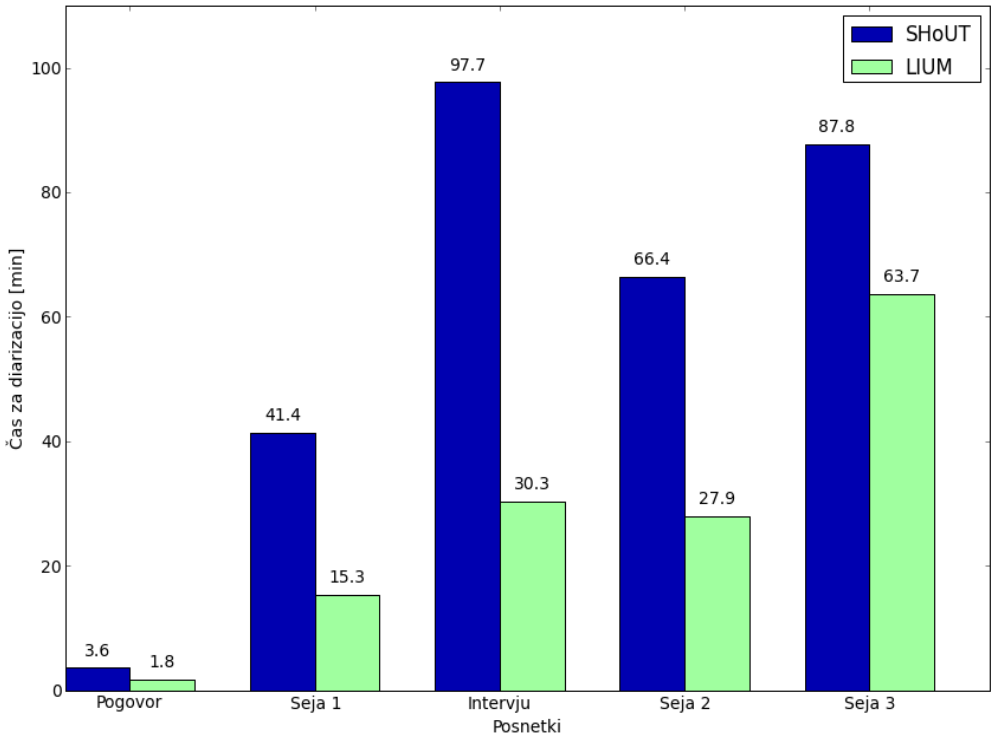
Posnetek	Pogovor		Seja 1		Intervju		Seja 2		Seja 3	
Orodje	SHoUT	LIUM	SHoUT	LIUM	SHoUT	LIUM	SHoUT	LIUM	SHoUT	LIUM
Čas diarizacije [s]	216,73	109,67	2483,97	917,55	5861,22	1816,55	3983,93	1674,57	5265,64	3822,61
Dolžina posnetka [s]	343,22		1842,30		3039,06		3147,65		4101,33	
Čas za sekundo posnetka	0,63	0,32	1,35	0,50	1,93	0,60	1,27	0,53	1,28	0,93
DER [%]	**33,12**	**47,41**	**7,34**	**3,48**	**6,04**	**5,86**	**17,47**	**21,12**	**14,99**	**52,96**
Referenčno št. govorcev	3		5		2		11		10	
Zaznano št. govorcev	2	6	8	8	3	5	9	10	10	7

Tabela 2: Rezultati diarizacije z orodjema SHoUT in LIUM.

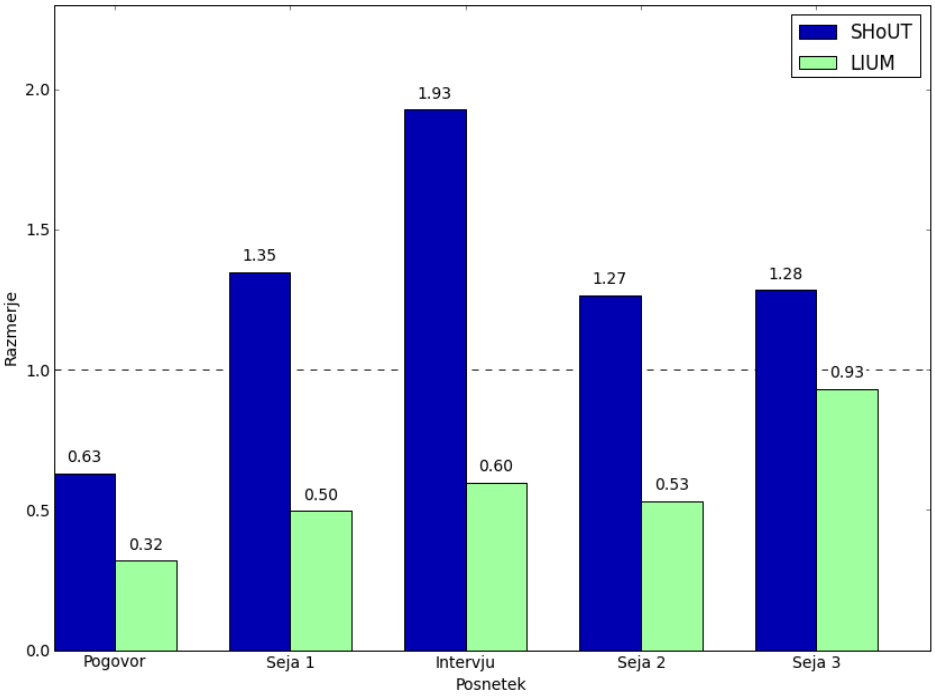


Ker so kriteriji omiljeni, rezultati niso neposredno primerljivi z rezultati, dobljenimi s strani avtorjev orodij. Obe orodji dosegata dobre rezultate za krajše posnetke, kjer ni veliko govorcev in kjer je posnetek narejen v nadzorovanem okolju. Kjer je govorcev več, so večje tudi napake. Še posebej je to očitno pri Seji 3, pri kateri je orodje LIUM združilo preveč razredov in tako zaznalo 7 govorcev namesto dejanskih 10, zaradi česar je potem tudi napaka močno narasla, na kar 53,0 odstotka.

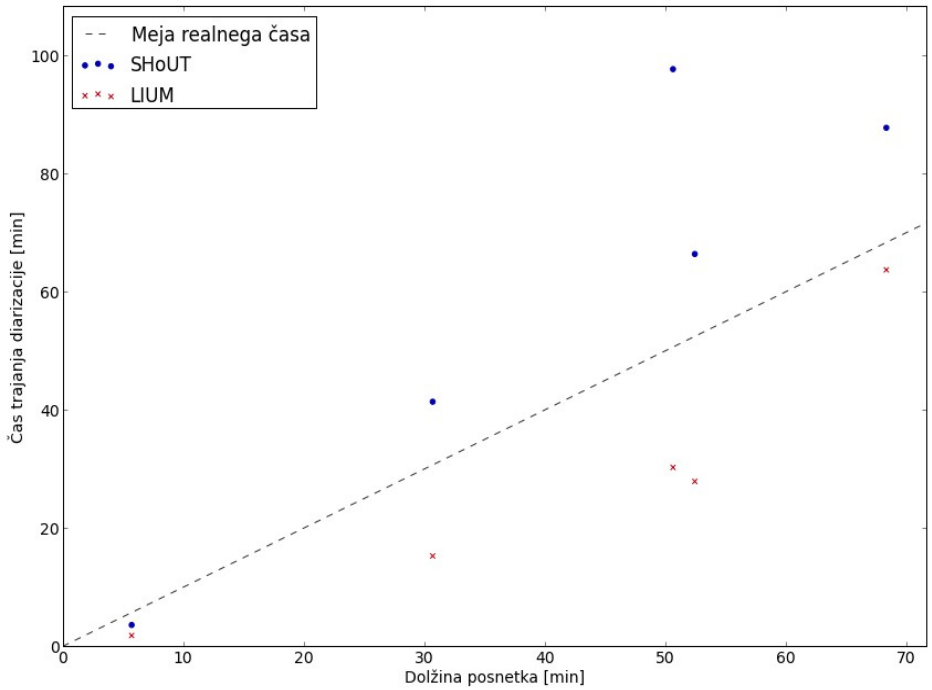
Orodje SHoUT je, čeprav je govor v slovenskem jeziku, konsistentno dosegalo dobre do zadovoljive rezultate, razen v primeru pogovora. Pri pogovoru je namreč kar 15,8 % časa posnetka takega, kjer se več govorcev prekriva, nobeno od orodij pa ne zaznava prekrivanja več govorcev.



Merili smo tudi čas, porabljen za diarizacijo. Ker so posnetki dolgi in ker je diarizacija procesno zahtevna, je tudi čas diarizacije dolg. Orodje LIUM je konsistentno opravilo diarizacijo hitreje od dolžine posnetka, medtem ko je orodje SHoUT za daljše posnetke vedno potrebovalo več od dolžine posnetka.



Slika 10: Razmerje med časom za diarizacijo in dolžino posnetka. Črtano je označena meja, pod katero je možno diarizacijo izvesti v realnem času.



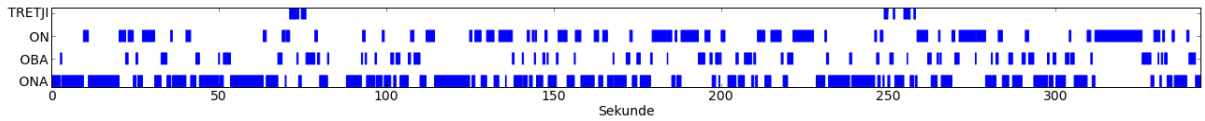
Iz grafov je razvidno, da orodje SHoUT porabi občutno več časa od orodja LIUM. Orodje SHoUT je zaradi svoje časovne zahtevnosti neprimerno za diarizacijo daljših posnetkov, če želimo diarizacijo opraviti v realnem času. Orodje LIUM je s tega vidika boljše, saj je konsistentno dosegalo čase, ki so bili manjši od dolžine posnetka.

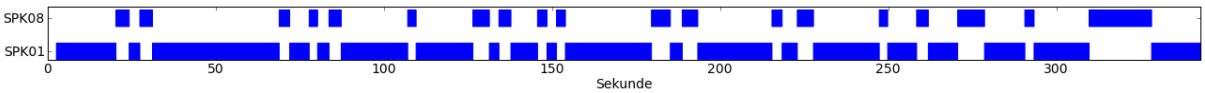
Pri intervjuju je pri orodju SHoUT prišlo do izraza veliko število prehodov govorca, zaradi katerega mora orodje preverjati in združevati veliko število nesosednjih razredov. Posledično čas diarizacije precej odstopa od drugih posnetkov, kljub temu pa je natančnost precej visoka. Pri orodju LIUM tovrstna časovna odvisnost od števila prehodov govorcev ni razvidna iz grafov.

5.1 Vizualizacija diarizacije

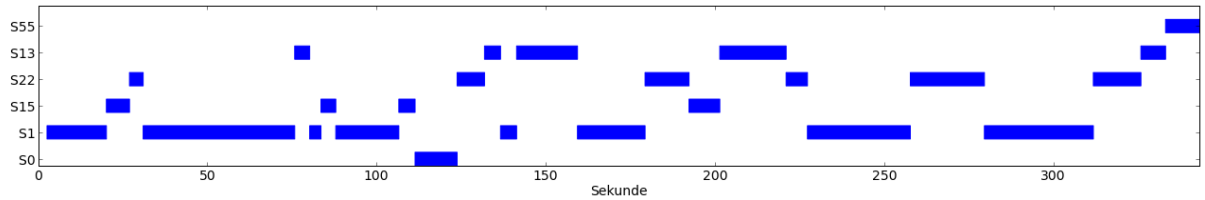
Za lažjo predstavitev smo diarizacije vizualizirali v obliki več vzporednih časovnic. Vsak tak graf s časovnicami predstavlja eno diarizacijo, bodisi referenčno, bodisi dobljeno z orodjem za diarizacijo. Na vodoravni osi je predstavljen čas z izhodiščem ob začetku posnetka, ob navpični osi pa so razporejeni govorci. Ob vsakem času je aktiven največ en govorec.

5.1.1 Diarizacija posnetka Pogovor





V pogovoru se v večini izmenjujeta dva govorca, označena z ON in ONA. Kadar govorita

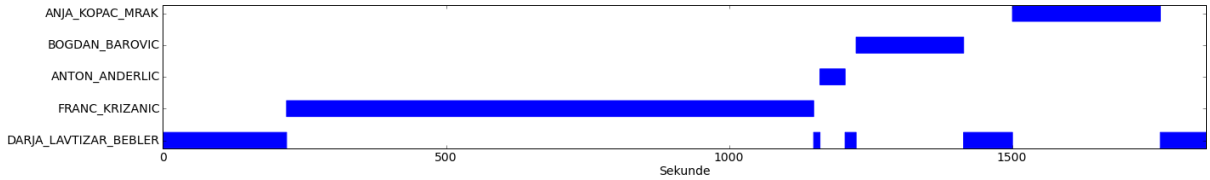


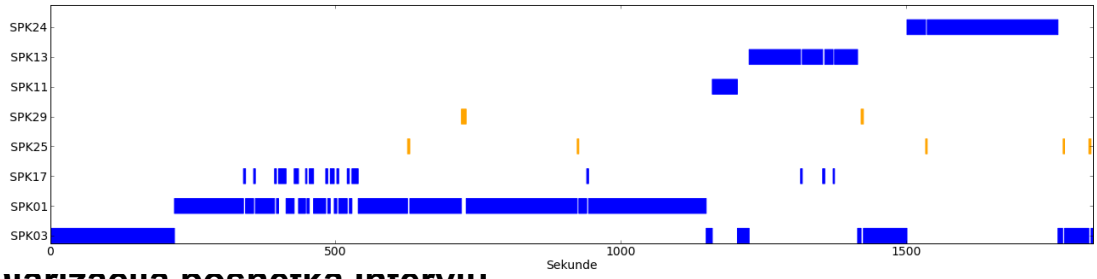
oba, je to označeno z OBA. Takega govora je kar za 15,8 %, zaradi česar je to tudi teoretična

minmalna napaka pri diarizaciji pogovora. Posnetek je bil sneman na prostem, zaradi česar so v ozadju slišni tudi različni šumi, na primer petje ptic.

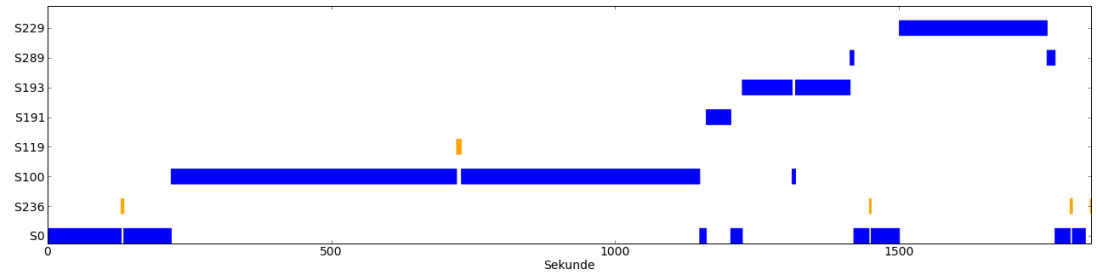
5.1.2 Diarizacija posnetka Seja 1

Pri posnetku Seja 1 je govorcev razmeroma malo. Obe orodji sta dosegli zelo dober rezultat. Napaka pri diarizaciji z orodjem SHoUT je 7,34 %, pri diarizaciji z orodjem LIUM pa 3,48 %. Pri diarizaciji z orodjem SHoUT je pri času okrog 500 sekund od začetka viden konsistenten vpliv šuma iz ozadja, ki ga je orodje zaznalo kot ločenega govorca. V tem konkretnem primeru gre za rožljanje skodelic in pribora. Z oranžno so označeni odseki, kjer je orodje zaznalo govorca neobstoječega govorca.





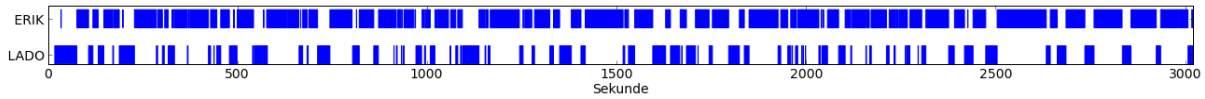
5.1.3 Diarizacija posnetka Intervju



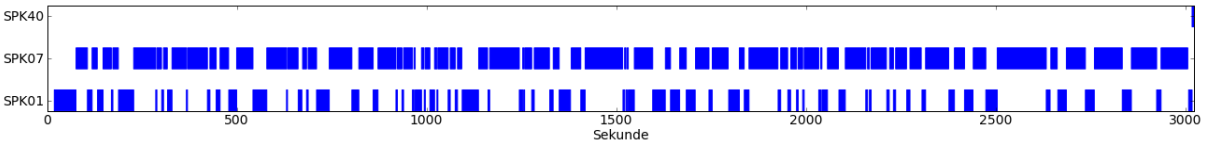
Posnetek Intervju je bil z vidika dosežene natančnosti orodij pri diarizaciji eden izmed boljših. S časovnic je jasno razvidno ujemanje diarizacij z referenčno.

Obe orodji sta zaznali uvodno in zaključno glasbo kot dodatne govorce. Orodje SHoUT ni zaznalo, da je ob koncu glasba in ne govor. Orodje LIUM je začetno in končno glasbo prepoznalo kot govor, ter celo, da gre za iste odseke glasbe.

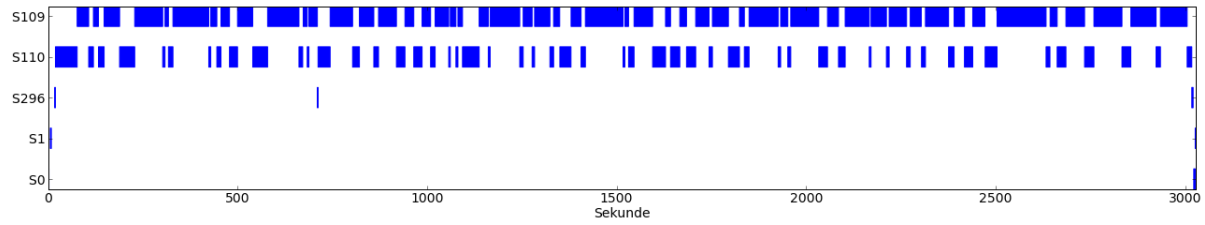
Čeprav je posnetek zelo dolg in vsebuje veliko interaktivnih prehodov med govorcema ter tudi nekaj prekrivanja govorcev, lahko iz dobljenega rezultata vidimo, da k natančni diarizaciji veliko pripomore tudi kvalitetna snemalna oprema in ozvočitev govorcev, zaradi česar je potem v posnetem govoru zelo malo motečih šumov.



Slika 18: Časovnica referenčne diarizacije posnetka Intervju



Slika 19: Časovnica diarizacije posnetka Intervju z orodjem SHoUT



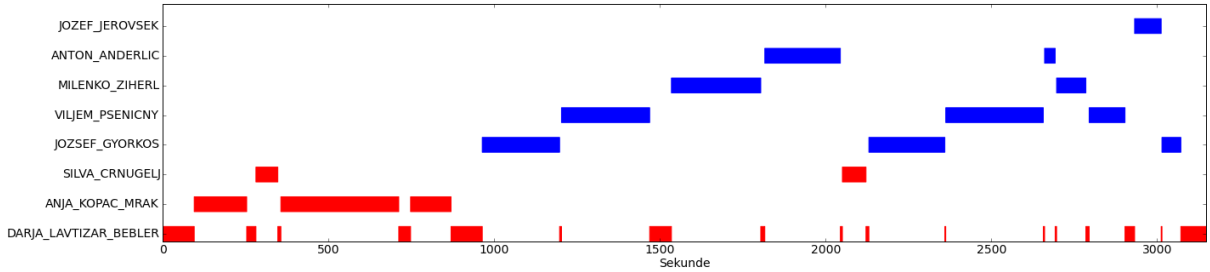
Slika 20: Časovnica diarizacije posnetka Intervju z orodjem LIUM

5.1.4 Diarizacija posnetka Seja 2

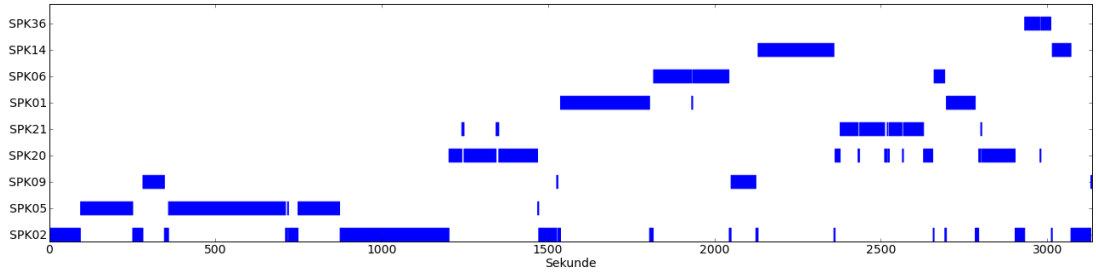
Pri posnetku Seja 2 je prišlo do kar nekaj napak, tako pri diarizaciji z orodjem SHoUT (17,47 % DER), kot tudi pri diarizaciji z orodjem LIUM (21,12 % DER). Obe orodji sta delali večje napake.

Tako je orodje SHoUT narobe v govorcu SPK02 združilo dva govorca iz referenčne diarizacije, vidno pa je tudi, ni združilo govorcev SPK20 in SPK21, ki sta v referenčni diarizaciji en govorec.

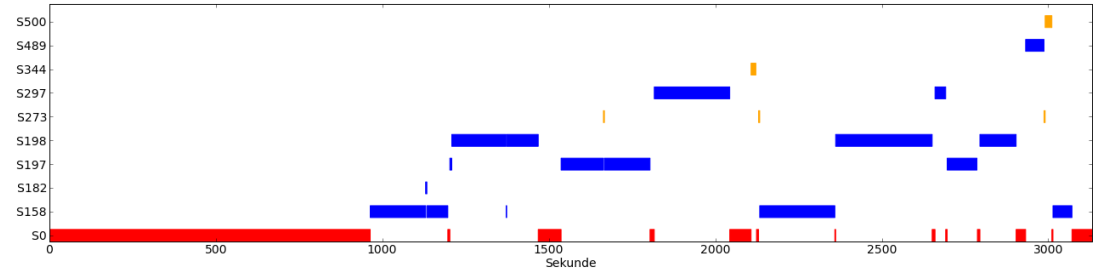
Orodje LIUM pa je vse govorce ženskega spola združilo v enega govorca.



Slika 21: Časovnica referenčne diarizacije posnetka Seja 2. Z rdečo so označeni govorci ženskega spola.



Slika 22: Časovnica diarizacije posnetka Seja 2 z orodjem SHoUT



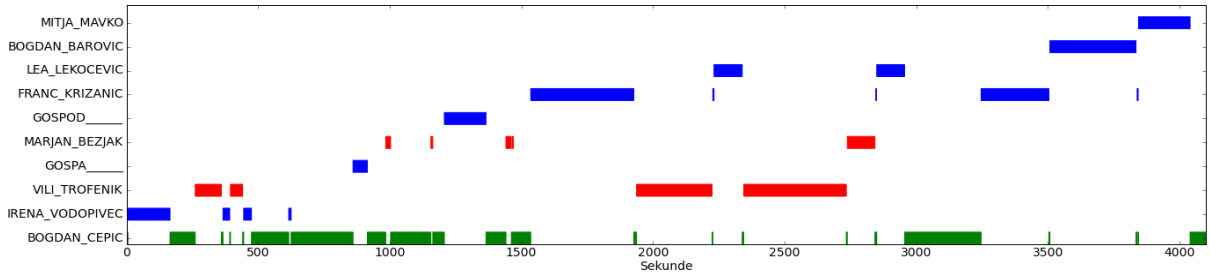
Slika 23: Časovnica diarizacije posnetka Seja 2 z orodjem LIUM. Z rdečo je označen govorec, v katerem so združeni vsi govorci ženskega spola. Z oranžno so označeni odvečni govorci.

5.1.5 Diarizacija posnetka Seja 3

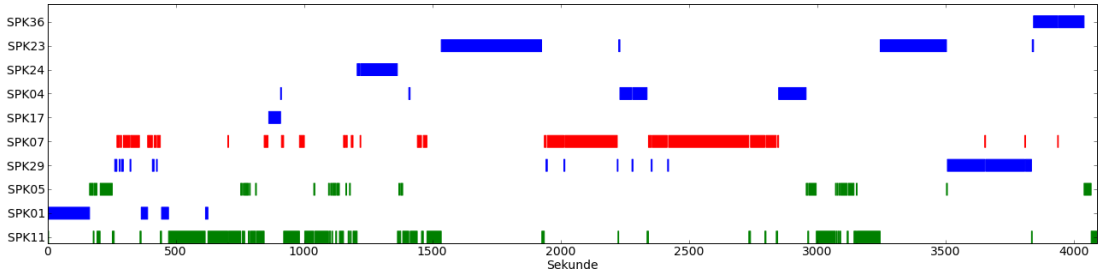
Pri diarizaciji posnetka Seja 3 orodje LIUM popolnoma odpovedalo, saj je dobljena diarizacija večino časa posnetka dodelila enemu govorcu, posledično pa je napaka pri

diarizaciji narasla na 52,96 %. Tako dobljena diarizacija očitno kaže na nekatere primere, v katerih orodje LIUM odpove.

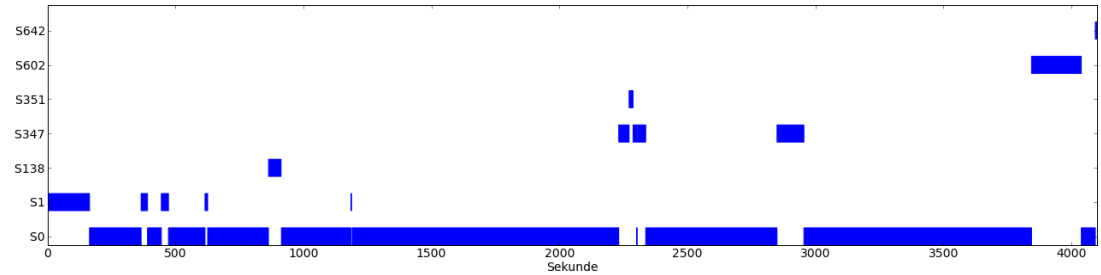
Diarizacija z orodjem SHoUT je s skupno napako diarizacije v višini 14,99 % podobno natančna kot pri posnetku Seja 2. Iz časovnic je razvidno, da je orodje SHoUT napačno združilo dva govorca, ki sta na časovnici označena rdeče, in premalo združilo dva govorca, kar je na časovnici označeno zeleno.



Slika 24: Časovnica referenčne diarizacije posnetka Seja 3



Slika 25: Časovnica diarizacije posnetka Seja 3 z orodjem SHoUT



Slika 26: Časovnica diarizacije posnetka Seja 3 z orodjem LIUM

6 Sklepne ugotovitve

Zanimivo je, da orodje SHoUT kljub temu, da na slovenskem govoru ni bilo naučeno, deluje razmeroma dobro, pri čemer je verjetno pomagalo razvojno vodilo, da se mora sistem čim manj zanašati na kakršno koli predznanje o posnetku.

Orodje LIUM je nekoliko manj uspešno in bi za doseganje boljše natančnosti pri diarizaciji slovenskega govora morali dodatno preučiti in nastaviti parametere. K temu morda še najbolj nakazuje združevanje ženskega govora pri diarizaciji posnetka Seja 2. Zanimivo je, da za doseganje dobrih rezultatov, kot je razvidno iz posnetka Seja 1, zadostuje tudi 13 koeficientov MFCC.

Vpliv opreme za snemanje je zelo dobro viden na rezultatih. Pri posnetku Intervju, kjer je bila uporabljena profesionalna snemalna oprema in sta bila govorca strokovno ozvočena, sta obe orodji dosegli dober in zelo izenačen rezultat.

Posnetek Pogovor se je za diarizacijo izkazal za najslabšega, kar je ob upoštevanju pogojev nastanka posnetka pričakovano. Iz petja ptic in šuma okolice lahko sklepamo, da je bil posnetek sneman na prostem. Za vse govorce je bil uporabljen isti mikrofon, zaradi česar je jakost govorca zelo odvisna od bližine mikrofona govorcu. Dodatno posnetek vsebuje zelo veliko prekrivanja govora, kar onemogoča enostavno določanje govorca in posledično navzdol omejuje napako diarizacije.

Pri posnetkih sej so rezultati slabši ali primerljivi z diarizacijo posnetka Intervju. Pri diarizaciji posnetka Seja 1, ki je sicer od vseh posnetkov sej najbolj natančno diariziran, je viden vpliv šuma iz prostora. Pri posnetkih daljših sej in sej z veliko govorci postane problem napačno združevanje. Pri velikem številu govorcev se namreč zaradi hitrega naraščanja števila primerjav zgodi, da imata dva govorca podobne govorne značilnosti. Ta pojav je znan tudi kot rojstnodnevni paradoks[3].

V prid dobri diarizaciji posnetkov sej gre način rabe mikrofonov. Vpliv lastnosti mikrofona na zaznavanje zvoka ni zanemarljiv, saj je natančnost prepoznave govora ob uporabi istega mikrofona, s katerim je bil akustični model naučen, bistveno boljša. Pri sejah se

3 Paradoks rojstnega dne opisuje verjetnost, da imata dve poljubni osebi v dani skupini rojstni dan istega dne v letu. Ker število kombinacij narašča z *N!*, verjetnost hitro narašča in že pri 23 ljudeh doseže 50%, kar ni intuitivno in marsikoga preseneti.

uporablja bližnji mikrofon z razmeroma malo šuma, pogosto pa govorci uporabljajo tudi ločene mikrofone, kar dodatno koristi k različnosti zaznavanja posameznega govorca, to pa koristi postopku diarizacije.

Za uporabo na slovenskem govoru je orodje SHoUT sprejemljivo in dosega nivoje napake, primerljive z uporabo orodij za diarizacijo na angleškem govoru, orodje LIUM pa ima z nekaterimi posnetki težave, zato bi bilo smiselno preučiti možnosti, da bi orodju podali dodatne, za slovenski govor ustrezne oziroma na slovenskem govoru naučene parametre.

Obema orodjema bi prav tako lahko podali tudi druge metapodatke. Za marsikateri posnetek bi na primer bilo možno pridobiti dobro oceno števila govorcev. Za seje je to možno pridobiti iz prepisa seje na strani Državnega zbora z razmeroma enostavnim procesiranjem besedila, pri čemer pa poleg števila govorcev dobimo tudi vrstni red govorcev. Iz obsega besedila bi lahko okvirno sklepali tudi na velikostni razred dolžine govora posameznega govorca. Pri intervjuju je prav tako vnaprej znano, da sta govorca le dva. Ta informacija bi pomagala, da orodje ne bi preveč zmanjšalo števila razredov, kakor se je zgodilo ob uporabi orodja LIUM na posnetku Seja 3.

Raziskave na področju diarizacije se vedno bolj premikajo v smer analize posnetkov slabše kakovosti, torej od studijskih posnetkov do posnetkov sej oziroma sestankov in končno do posnetkov govora v okolju z veliko šuma. Ti spremenjeni pogoji zahtevajo nove dodatne stopnje procesiranja in tako se že delajo evalvacije, ki se zanašajo na dodatne vire informacij, dobljenih z uporabo polja mikrofonov in video zapisa posnetka.

Zaradi statistične narave uporabljenih algoritmov in nepredvidljivosti vhodnih podatkov predstavlja diarizacija zelo zahteven problem in ni za pričakovati, da bi z manjšo količino učnih posnetkov lahko dobili algoritme, ki bi delovali idealno tudi na nepričakovanih vhodnih podatkih. Kljub temu pa orodje SHoUT dokazuje, da je možno dobiti neidealne, a stabilne algoritme, ki delujejo tudi na nepredvidenih vhodnih podatkih, kakršen je na primer orodju tuj jezik.

7 Literatura

[1] Friedland, G., Janin, A., Imseng, D., Anguera Miro, X., Gottlieb, L., Huijbregts, M., Knox, M.T., Vinyals, O., "The ICSI RT-09 Speaker Diarization System," *Audio, Speech, and Language Processing, IEEE Transactions on* , vol.20, no.2, 371-381, februar 2012.

[2] (2013) L'Association Francophone de la Communication Parlée, *Campagne d'évaluation des Systèmes de Transcription Enrichie d'Émissions Radiophoniques*. Dostopno na http://www.afcp-parole.org/camp_eval_systemes_transcription/

[3] A. Žgank, D. Verdonik, Z. Kačič, "Slovenska baza BNSI broadcast news za razpoznavanje tekočega govora", *Elektrotehniški vestnik* 75 (3): 85-90, 2008.

[4] Rabiner, Lawrence R., Schafer, Ronald W., "Digital processing of speech signals", Prentice-Hall, 1978, pogl. 7.

[5] C. M. Bishop, "Pattern Recognition and Machine Learning", Springer, 2006.

[6] Huijbregts, Marijn, "Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled", PrintPartners Ipskamp, 2008. Doktorska disertacija.

[7] (2013) Information Technology Laboratory, *Rich Transcription Spring 2006 Evaluation.* Dostopno na http://www.itl.nist.gov/iad/mig//tests/rt/2006-spring/

[8] S. Meignier, T. Merlin, "LIUM SpkDiarization: An Open Source Toolkit For Diarization," v Proc. CMU SPUD Workshop, Marec 2010, Dallas (Texas, USA). Orodje dostopno na http://lium3.univ-lemans.fr/diarization/doku.php/download